ABSTRACT

The present invention involves a system and method that facilitate extracting data from messages for spam filtering. The extracted data can be in the form of features, which can be employed in connection with machine learning systems to build improved filters. Data associated with the subject line, timestamps, and the message body can be extracted and employed to generate one or more features. In particular, subject lines and message bodies can be examined for consecutive, repeating characters, blobs, the association or distance between such characters, blobs and non-blob portions of the message. The values or counts obtained can be broken down into one or more ranges corresponding to a degree of spaminess. Presence and type of attachments to messages, percentage of non-white-space and non-numeric characters of a message, and determining message delivery times can be used to identify spam. A time-based delta can be computed to facilitate determining the delivery time.